Transcrição

Começaremos o novo tópico de **Medidas de Tendência Central**.

Neste grupo de assuntos, falaremos sobre a média aritmética, mediana e média. Ao final, abordaremos a relação entre essas medidas.

Faremos um `DataFrame` dentro da variável `df` que nos ajudará no entendimento desses cálculos. Em seguida, o aplicaremos em nosso dataset mai

Este será relativo às notas de três alunos fictícios em sete matérias escolares diferentes.

```
df = pd.DataFrame(data = {'Fulano': [8, 10, 4, 8, 6, 10, 8],
                          'Beltrano': [10, 2, 0.5, 1, 3, 9.5, 10],
                          'Sicrano': [7.5, 8, 7, 8, 8, 8.5, 7]},
                  index = ['Matemática',
                           'Português',
                           'Inglês',
                           'Geografia',
                           'História',
                           'Física',
                           'Química'])
df.rename_axis('Matérias', axis = 'columns', inplace = True)
df
```

Matérias	Beltrano	Fulano	Sicrano
Matemática	10.0	8	7.5
Português	2.0	10	8.0
Inglês	0.5	4	7.0
Geografia	1.0	8	8.0
História	3.0	6	8.0
Física	9.5	10	8.5
Química	10.0	8	7.0

Nas etapas anteriores, fizemos uma sumarização dos dados, reduzindo-os para tentarmos entender um conjunto que é bastante grande, como em noss

Tentaremos sumarizar ainda mais por meio de uma medida que pegará uma informação importante do conjunto.

Começaremos falando sobre a **média aritmética**; é basicamente o centro de massa da distribuição de uma variável, equilibrando-a. Por ser muito se variáveis, justamente por conta desta sensibilidade.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Em nossa variável `Renda` por exemplo, vimos que a grande maioria das pessoas recebe um baixo rendimento mensal e poucas possuem altas rendas representar este conjunto de dados corretamente.

De volta ao `DataFrame` deste passo, calcularemos manualmente somente a média neste primeiro momento, e depois lidaremos com as demais medi

Na primeira célula da parte "3.1 Média aritmética", copiaremos as notas de `Fulano` e as colaremos dentro de parênteses para aplicarmos a fórmula.

Logo, trocaremos as vírgulas pelo símbolo de `+` para realizarmos a somatória e dividiremos o resultado por `7`.

```
(8 + 10 + 4 + 8 + 6 + 10 + 8) / 7
```

Como retorno, veremos a média `7.714285714285714` exibida na saída.

Porém, não poderemos calcular assim todas as vezes, pois em nosso dataset existem mais de 70 mil registros por exemplo, o que tornaria a operação

Começaremos com a variável `df[]` com `'Fulano'`, pois a biblioteca Pandas disponibiliza uma função **`.mean()`** para calcular a média.

```
df['Fulano'].mean()
```

O retorno será o mesmo valor apresentado antes, porém fizemos a mesma operação de maneira mais simples.

De volta ao nosso dataset oficial, calcularemos da mesma forma para descobrirmos a média da `Renda`.

```
dados.Renda.mean()
```

Com a saída do valor `2000.3831988547631`, veremos que a média dos rendimentos dentro da pesquisa é de aproximadamente R$2.000,00.

Mais adiante neste curso, veremos as influências dos extremos neste cálculo para a coerência da análise.

Em uma nova célula, veremos novamente só os cinco primeiros registros de nosso dataset.

```
dados.head()
```

	UF	Sexo	Idade	Cor	Anos de Estudo	Renda	Altura
0	11	0	23	8	12	800	1.603808
1	11	1	23	2	12	1150	1.739790
2	11	1	35	8	15	880	1.760444
3	11	0	46	2	6	3500	1.783158
4	11	1	47	8	9	150	1.690631

No começo do curso, abordamos os tipos de variáveis e vimos que alguns deles não permitem o cálculo da média por motivos óbvios, como no caso

Poderemos usar estas variáveis para nos auxiliar nas análises como filtros ou *by*, como a renda média por cada sexo por exemplo.

Apagaremos a linha de `head()` e chamaremos nossos `dados` novamente. Em seguida, aplicaremos a funcionalidade **`.groupby()`** do Pandas para a

```
dados.groupby(['Sexo']).mean()
```

Sexo	UF	Idade	Cor	Anos de Estudo	Renda	Altura
0	31.901991	44.046554	5.038685	9.120169	2192.441596	1.603808
1	31.937728	44.127554	5.018906	10.258584	1566.847393	1.699670

Com esta execução, o sistema calculará a média de todas as variáveis do dataset para cada sexo, incluindo aquelas que não poderiam ser calculadas.

Portanto, escolheremos quais queremos calcular. Antes de `.mean()`, pediremos somente a `'Renda'`.

```
dados.groupby(['Sexo'])['Renda'].mean()
```

Sexo	
0	2192.441596
1	1566.847393

Com isso, conseguiremos entender melhor o funcionamento da média aritmética e como aplicá-la.

A seguir, partiremos para a **mediana**.